AMENDMENTS TO THE SPECIFICATION

Please amend the Title as follows:

~~DATABASES OF~~ LIBRARIES OF REGULATORY SEQUENCES, METHODS OF MAKING AND USING SAME

Please amend the paragraph beginning on page 25, line 22 as follows:

Techniques for determining nucleic acid and amino acid sequence similarity are known in the art. Typically, such techniques include determining the nucleotide sequence of, *e.g.*, an accessible region of cellular chromatin, and comparing these sequences to a second nucleotide sequence. Genomic sequences can also be determined and compared in this fashion. In general, "identity" refers to an exact nucleotide-to-nucleotide or amino acid-to-amino acid correspondence of two polynucleotides or polypeptide sequences, respectively. Two or more sequences (polynucleotide or amino acid) can be compared by determining their "percent identity." The percent identity of two sequences, whether nucleic acid or amino acid sequences, is the number of exact matches between two aligned sequences divided by the length of the shorter sequence and multiplied by 100. An approximate alignment for nucleic acid sequences is provided by the local homology algorithm of Smith and Waterman, Advances in Applied Mathematics 2:482-489 (1981). This algorithm can be applied to amino acid sequences by using the scoring matrix developed by Dayhoff, Atlas of Protein Sequences and Structure, M.O. Dayhoff ed., 5 suppl. 3:353-358, National Biomedical Research Foundation, Washington, D.C., USA, and normalized by Gribskov, Nucl. Acids Res. 14(6):6745-6763 (1986). An exemplary implementation of this algorithm to determine percent identity of a sequence is provided by the Genetics Computer Group (Madison, WI) in the "BestFit" utility application. The default parameters for this method are described, for example, in the Wisconsin Sequence Analysis Package Program Manual, Version 8 (1995) (available from Genetics Computer Group, Madison, WI). An additional method of establishing percent identity in the context of the present disclosure is to use the MPSRCH package of programs copyrighted by the University of Edinburgh, developed by John F. Collins and Shane S. Sturrok, and distributed by IntelliGenetics, Inc. (Mountain View, CA). From this suite of packages the Smith-Waterman algorithm can be employed where default parameters are used for the scoring table (for example, gap open penalty of 12, gap extension penalty of one, and a gap of six). From the data generated the "Match" value reflects "sequence identity." Other suitable programs for calculating the percent identity or similarity between sequences are generally known in the art, for example, another alignment program is BLAST, used with default parameters. For example, BLASTN

and BLASTP can be used using the following default parameters: genetic code = standard; filter = none; strand = both; cutoff = 60; expect = 10; Matrix = BLOSUM62; Descriptions = 50 sequences; sort by = HIGH SCORE; Databases = non-redundant, ~~GenBank~~GENBANK™ + EMBL + DDBJ + PDB + ~~GenBank~~GENBANK™ CDS translations + Swiss protein + Spupdate + PIR. Details of these programs can be found, for example, ~~at the following~~ on the Internet ~~internet address: http://www.ncbi.nlm.gov, accessed on October 22, 2001~~. When claiming sequences relative to sequences described herein, the range of desired degrees of sequence identity is approximately 80% to 100% and any integer value therebetween. Typically the percent identities between the disclosed sequences and the claimed sequences are at least 70-75%, preferably 80-82%, more preferably 85-90%, even more preferably 92%, still more preferably 95%, and most preferably 98% sequence identity to the reference sequence.

Please amend the paragraph beginning on page 86, line 17 as follows:

Signal transduction pathways mediate gene expression through a membrane receptor capable of transducing signal to the nucleus through an intricate network of molecules, which in turn control gene expression. Various signal transduction pathways are currently under study. *See, e.g.,* (1999) *Science* 284:755-770 and articles cited therein and on the Internet ~~address www.stke.org, accessed on August 11, 2000~~.

Please amend the paragraph beginning on page 92, line 1 as follows:

The sequence(s) to be compared against a comparison sequence is(are) typically obtained from an internal database populated as set forth supra, but can also be obtained from an external database. In general, an external database refers to a database that is located outside of the internal database. Most typically, such a database is one that has not been developed and maintained by the entity conducting the comparison but rather has been developed by an entity other than the one that maintains the internal database. Examples of external databases include ~~GenBank~~GENBANK™ and other associated databases that are maintained by the National Center for Biotechnology Information (NCBI), part of the National Library of Medicine. Other examples of external databases include the Blocks database maintained by the Fred Hutchinson Cancer Research Center in Seattle, WA, and the Swiss-Prot site maintained by the University of Geneva. The comparison or reference sequences can be stored with the sequences being compared on the internal database or can be stored in a separate database that is either another internal database or an external database.

Please amend the paragraph beginning on page 96, line 29 as follows:

An exemplary basic networked system suitable for conducting the sequence analyses described herein is depicted in FIG. 17 which is a block diagram showing the general configuration of one example of a suitable system. As shown in FIG. 17, the networked system 70 includes a workstation 80, an internal database 82 located within an organization, an optional external database 84 typically located outside the organization, a communications modem 86 and a network system 88 that allows the workstation 80 to access information from external data storage systems such as the external database 84. Thus, for example, the workstation 80 can be connected via the modem 86 and network system 88 to another database 84 at a research institute or university. As indicated supra, in some instances the external database 84 is the ~~GenBank~~GENBANK™ database or a similar type database maintained by a research institute, university or company. In this manner, sequence information stored on the internal database 82 can be supplemented with sequence information and other related types of information concerning the sequences from an external database 84.

Please amend the paragraph beginning on page 99, line 29 as follows:

However, for systems 100 that do utilize a World Wide Web server and clients, the system should support a TCP/IP protocol. Local networks such as this are sometimes referred to as "Intranets." Such intranets have the advantage that they permit facile communication with public domain databases on the World Wide Web, such as ~~GenBank~~GENBANK™. Hence, in certain systems, the auxiliary computers 104a, 104b can directly access data (e.g., via Hypertext links) residing on the Internet databases using a HTML interface provided by Web browsers and Web server 118.

Please amend the paragraph beginning on page 102, line 21 as follows:

The hit source field indicates the database source from which the comparison or reference sequence was obtained. Thus, for example, the hit source field can indicate whether the comparison sequence was obtained from an internal or an external database such as a public domain database (e.g., ~~GenBank~~GENBANK™). In the case of a public domain database such as ~~GenBank~~GENBANK™, the field can specifically identify the database within ~~GenBank~~GENBANK™ from which the sequence was obtained. A hit description field can provide descriptive information regarding the comparison or reference sequence. This information can be provided by the user or, in the case of an external database maintained by another organization, the information can simply come from the information provided by the organization maintaining the external database.

Please amend the paragraph beginning on page 103, line 22 as follows:

The database can optionally include a table denoted as an External Hit Table 156 to summarize information on hits against sequences stored in public domain sequence databases such as ~~GenBank~~GENBANK™, for example. Hence, if a sequence in the Sequence Project Table 152 matches a sequence in the public database with the requisite degree of specificity as input by the user, then the match from the public database is provided as a record in the External Hit table 156. Typically, each record in this table includes at least a hit ID and a hit description to fully identify the sequence. In like manner, the database can include an Internal Hit Table 158 to summarize information on hits against sequences stored on an internal database. This can be useful when different collections of accessible regions are stored on one or more storage devices within an organization.

Please amend the paragraph beginning on page 111, line 12 as follows:

Examination of the nucleotide sequence of the human TEF-3 (transcription enhancer factor-3, also known as RTEF-1) gene (~~GenBank~~GENBANK™ accession Number AC005911) between –2,940 and +3,060, with respect to the P1 transcription startsite, reveals the presence of a CpG-rich region between -660 and +840, marked by the presence of 30 Hpa II sites (*i.e.*, a CpG island). This sequence was searched for the presence of the sequence 5'-TTAA-3' which is the recognition site for the restriction enzyme Mse I. The search revealed the existence of 13 Mse I sites in this region. Of the 14 predicted Mse I fragments from this region, 13 were smaller than 900 nucleotide pairs. The remaining fragment was predicted to have a length of 1,935 nucleotide pairs, extending from –992 to +943, and contained all 30 of the Hpa II sites in the –660 to +840 CpG island. *See* Figure 5. This large, 1,935 nucleotide pair fragment is easily separable, by gel electrophoresis or other size separation methods, from all other products of Mse I digestion of this region.

Please amend the paragraph beginning on page 111, line 33 as follows:

An analysis similar to that described in example 7, *supra*, was conducted on an approximately 7-kilobase pair (kbp) segment of the human TRAF-3 (TNF Receptor-Associated Factor, also known as CAP-1) gene; ~~GenBank~~GENBANK™ Accession Number AF110907. In this case, a CpG-island, containing 38 Hpa II sites, is present between -840 and +900, with respect to the P2 transcriptional startsite. Analysis of the predicted sizes of Mse I fragments in this 7-kbp region revealed the existence of two large Mse I fragments, of 2,784 and 1,623 nucleotide pairs. All other Mse I fragments (nine in total) had sizes less than 800 nucleotide

pairs. *See* Figure 6. The larger of these fragments extends from -1,718 to +1,066, encompassing the CpG island.

Please amend the paragraph beginning on page 127, line 24 as follows:

In this example, chromatin immunoprecipitation (ChIP) was used to enrich a population of DNA fragments comprising regulatory sequences for the p16 tumor suppressor gene, by virtue of their association with acetylated histone H3. A CpG island is located in the p16 gene (~~GenBank~~GENBANK™ Accession No. AF022809), between about 30 nucleotide pairs upstream, and about 590 nucleotide pairs downstream of the transcriptional startsite identified by Hara et al. (1996) *Mol. Cell. Biol.* **16**:859-867. One form of regulation of genes associated with CpG islands is through methylation of C residues within the CpG island. Methylation is generally correlated with repression of transcription, while demethylation of methylated sequences can lead to transcriptional activation. In HCT15 cells, the p16 CpG island is methylated and the p16 gene is inactive. Treatment of HCT15 cells with 5-azacytidine (an inhibitor of CpG methylation) results in activation of p16 transcription.

Please amend the Abstract as follows:

Methods and compositions for the identification, isolation and characterization of regulatory DNA sequences in a cell of interest are provided. In particular, ~~Also provided are~~ libraries of regulatory sequences are provided, in which each member of the library comprises a polynucleotide comprising sequences from an accessible region of cellular chromatin ~~obtained according to the methods, and databases comprising collections of regulatory sequences for a particular cell of interest. In addition, various uses for the regulatory sequences so obtained, and uses for the databases of regulatory sequences, are provided. Also disclosed are computer systems and computer program products for utilizing the databases to conduct various genetic analyses, and uses of accessible regulatory sequences in the design of vectors bearing transgenes~~.

A clean copy of the Abstract is attached hereto as a separate sheet.